Predictive Analysis of Road Accidents Using Support Vector Machine, Decision Tree, and Random Forest

K.Pavani[1],, A.Anjali[2]

#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology Vijayawada

#2 Student in the Department of MCA, SRK Institute of TechnologyVijayawada

Abstract: Road accidents are increasing rapidly due to factors such as road conditions, environmental influences, and driver behavior. This paper proposes a hybrid machine learning model that integrates data preprocessing, feature selection, clustering, and classification techniques to predict accident occurrences. Algorithms such as Support Vector Machine (SVM), Decision Tree, and Random Forest are applied to analyze historical accident data and improve prediction accuracy. The proposed system effectively identifies accident-prone scenarios and provides better decision support for reducing road accidents.

Index terms - *— Road Accident Prediction, Machine Learning, Data Mining, Support Vector Machine (SVM), Decision Tree, Random Forest, Clustering, Feature Selection, Data Preprocessing, Predictive Analytics*

1. INTRODUCTION

Road accidents have become a major global concern, especially in countries like India where the number of vehicles is increasing rapidly. Factors such as over-speeding, poor road conditions, environmental influences like weather, and human negligence significantly contribute to accident occurrences. According to recent studies, a large number of fatalities are reported every year due to road accidents, making it essential to develop intelligent systems that can predict and prevent such incidents.

With the advancement of Artificial Intelligence and Machine Learning, data-driven approaches are widely used to analyze large volumes of accident data and identify hidden patterns. Machine learning techniques enable the extraction of meaningful insights from historical data, helping in predicting the likelihood of accidents in specific regions and conditions. These predictions can assist government authorities and road users in taking preventive measures to reduce accident risks.

In this paper, a hybrid machine learning model is proposed that combines data preprocessing, feature selection, clustering, and classification techniques to improve prediction accuracy. Algorithms such as Support Vector Machine (SVM), Decision Tree, and Random Forest are utilized to analyze accident-related data and generate reliable predictions. The proposed system aims to provide an efficient and scalable

solution for identifying accident-prone areas and enhancing road safety.

2. LITERATURE SURVEY

a) Discovering recurring anomalies in text reports regarding complex space systems:

A substantial amount of historical maintenance and problem databases are kept in unstructured text formats in many of the complex space systems that are now in use. Finding recurrent abnormalities and connections between problem reports that can point to more serious systemic issues is the issue we tackle in this research. We use data from discrepancy reports on software abnormalities in the Space Shuttle to demonstrate our methods. The focus and language of these free text reports fluctuate greatly since they are produced by a variety of individuals. We evaluate four common automated techniques for text anomaly detection found in the present text mining literature. We begin by describing the application of the Gaussian mixture model, sometimes known as k-means, to the term-document matrix. The second technique is the Sammon nonlinear map, which projects high-dimensional document vectors into two dimensions for grouping and visualization. The third approach, which depicts each document as a point on a high-dimensional sphere, is based on an investigation of the outcomes of using a clustering method, expectation maximization on a combination of von Mises Fisher distributions. We use clustering in this area to find sets of related texts. The findings come from a novel technique called spectral clustering, which embeds vectors from the term-document matrix in a high-dimensional space for grouping. Recommendations for the creation of an operational text mining system for the analysis of problem reports resulting from complicated space systems are included in the paper's conclusion. We also show the areas where this system has to be tailored for the space domain by contrasting such systems with general-purpose text mining methods.

b) Discovering recurring anomalies in text reports regarding complex space systems:

A substantial amount of historical maintenance and problem databases are kept in unstructured text formats in many of the complex space systems that are now in use. Finding recurrent abnormalities and connections between problem reports that can point to more serious systemic issues is the issue we tackle in this research. We use data from discrepancy reports on software abnormalities in the Space Shuttle to demonstrate our methods. The focus and language of these free text reports fluctuate greatly since they are produced by a variety of individuals. We evaluate four common automated techniques for text anomaly detection found in the present text mining literature. We begin by describing the application of the Gaussian mixture model, sometimes known as k-means, to the term-document matrix. The second technique is the Sammon nonlinear map, which projects high-dimensional document vectors into two dimensions for grouping and visualization. The third approach, which depicts each document as a point on a high-dimensional sphere, is based on an investigation of the outcomes of using a clustering method, expectation maximization on a combination of von Mises Fisher distributions. We use clustering in this area to find sets of related texts. The findings come from a novel technique called spectral clustering, which embeds vectors from the term-document matrix in a high-dimensional space for grouping. Recommendations for the creation of an operational text mining system for the analysis of problem reports

resulting from complicated space systems are included in the paper's conclusion. We also show the areas where this system has to be tailored for the space domain by contrasting such systems with general-purpose text mining methods.

c) Identifying the Stances of Topic Persons Using a Model-based Expectation-Maximization Method:

Finding similar-minded individuals in topic papers with conflicting opinions can aid readers in understanding the context of a subject and make topic reading easier. In this study, we provide an unsupervised technique to find subject individuals who have the same viewpoint. In particular, we cluster people into positively linked groups using a model-based Expectation-Maximization (EM) technique. Furthermore, we employ a weighted correlation coefficient and an off-topic block removal approach to eliminate off-topic text blocks and mitigate the text sparseness issue. Additionally, we offer an efficient initialization approach that produces suitable EM initializations. The results of our experiment show that the suggested strategy outperforms several popular clustering techniques and accurately clusters topic individuals with the same perspective. Additionally, the initialization technique produces stance identification results that are reliable and accurate.

d) ext Mining Analysis of Railroad Accident Investigation Reports:

Major train incidents are reported by the Transportation Safety Board of Canada and the National Transportation Safety Board of the United States. To find recurrent themes in significant train incidents, the text from these accident reports was examined using the text mining techniques of probabilistic topic modeling and k-means clustering. These studies' results show that the train accidents may be effectively categorized into many subjects. Additionally, the result indicates that grade crossing accidents, wheel flaws, track faults, and switching accidents are common accident categories. The discovery that bridge-related incidents are more common in Canadian reports is one of the main differences between the U.S. and Canadian reports. ASME Country-Specific Mortality and Growth Failure in Infancy and Young Children and Association With Material Stature, Copyright © 2016 View and sort data on baby and early childhood mortality, growth failure, and their correlation with maternal health using interactive visuals and maps.

e) Analysis of road accidents in India using data mining classification algorithms:

Classification is a model-finding procedure that divides the data into several classes according to certain limitations. The road accident data set from India is analyzed in this paper utilizing a variety of classification techniques, including logistic regression, linear regression, decision trees, SVM, Naïve Bayes, KNN, Random Forest, and gradient boosting algorithms. The performance metrics that are employed include execution time, accuracy, and error rate. The R data mining program is used for this investigation. KNN outperforms other algorithms in terms of performance.

3. METHODOLOGY

i) Proposed Work:

The proposed system focuses on developing an intelligent road accident prediction model using a hybrid machine learning approach. Initially, raw accident data collected from various sources is

subjected to data preprocessing techniques such as data cleaning, handling missing values, and normalization to improve data quality. Feature selection is then applied to extract only the most relevant attributes influencing accident occurrences, such as road conditions, weather factors, and traffic parameters. This step helps in reducing complexity and improving the efficiency of the model.

After preprocessing, clustering techniques are used to group similar accident patterns, which helps in understanding hidden relationships within the data. The clustered data is then passed through multiple machine learning classification algorithms such as Support Vector Machine (SVM), Decision Tree, and Random Forest. These models are trained and evaluated to determine the most accurate prediction results. The system selects the best-performing algorithm to predict the probability of accidents, thereby enabling early identification of accident-prone situations and supporting preventive decision-making.

ii) System Architecture:

The system architecture begins with the user interacting with the application interface, where an authentication check is performed to ensure secure access. If the user is unauthorized, access is denied; otherwise, the authorized user proceeds to upload the road accident dataset. Once the dataset is uploaded, the system allows the user to view and verify the data, ensuring that the input is correct and suitable for further processing. This initial stage ensures data integrity and secure system usage.

After data validation, the system executes machine learning algorithms on the processed dataset. The algorithms analyze the data and generate predictions regarding accident occurrences. The predicted results are then displayed to the user along with graphical visualizations for better understanding and analysis. Finally, the process concludes after presenting the output, enabling users to interpret accident trends and make informed decisions based on the generated insights.

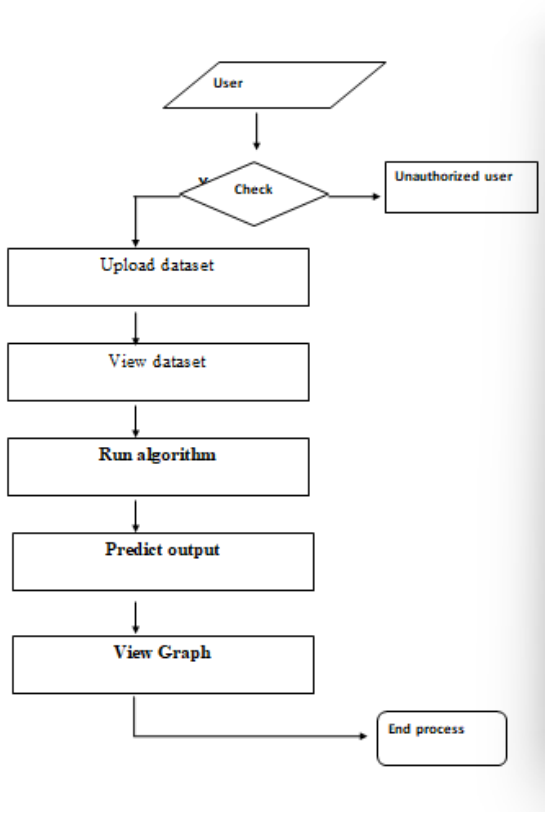

Fig1 proposed architecture

iii) Modules:

1. Data Collection Module

- Collects historical road accident data from datasets
- Includes features like weather, road condition, location, etc.

2. Data Preprocessing Module

- Cleans data by removing null and inconsistent values

- Normalizes data to improve model performance

3. Feature Selection Module

- Selects important attributes affecting accidents
- Reduces dimensionality and improves accuracy

4. Clustering Module

- Groups similar accident patterns using clustering techniques
- Helps identify hidden relationships in data

5. Machine Learning Module

- Applies SVM, Decision Tree, and Random Forest algorithms
- Trains models and compares performance

6. Prediction Module

- Predicts the probability of accident occurrence
- Selects the model with highest accuracy

7. Visualization Module

- Displays results using graphs and charts
- Helps users understand accident trends easily

iv) ALGORITHMS:

1. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification tasks. It works by finding an optimal hyperplane that separates data points into different classes with maximum margin. In this project, SVM is used to classify accident and non-accident scenarios based on input features such as road conditions and environmental factors. It performs well with high-dimensional data and provides accurate predictions.

2. Decision Tree

Decision Tree is a supervised learning algorithm that uses a tree-like structure to make decisions based on feature values. Each internal node represents a condition, and each branch represents an outcome leading to a final decision. In this system, Decision Tree helps in identifying the most influential factors contributing to road accidents and provides easy interpretability of results.

3. Random Forest

Random Forest is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It builds several trees using random subsets of data and features, and the final prediction is made based on majority voting. In this project, Random Forest enhances the reliability and robustness of accident prediction by producing more stable and accurate results.

4. Clustering Algorithm

Clustering is an unsupervised learning technique used to group similar data points based on their characteristics. In this model, clustering is applied before classification to identify patterns and similarities in accident data. This helps in improving the performance of classification algorithms by organizing the data into meaningful groups.

4. EXPERIMENTAL RESULTS

The proposed road accident prediction model was evaluated using real-world accident datasets after performing data preprocessing and feature selection. Multiple machine learning algorithms, including Support Vector Machine (SVM), Decision Tree, and Random Forest, were implemented and compared based on their prediction accuracy. The results indicate that the hybrid approach, which combines clustering with classification techniques, improves the overall

performance of the system by effectively identifying hidden patterns in the data.

Among the applied algorithms, Random Forest achieved the highest accuracy due to its ensemble learning capability, followed by SVM and Decision Tree. The system successfully predicted accident-prone scenarios and provided graphical visualizations for better interpretation of results. These findings demonstrate that the proposed model is efficient, reliable, and suitable for real-time accident prediction and decision-making support systems.

Accuracy: A test's accuracy is determined by its capacity to distinguish between healthy and ill cases. To gauge the accuracy of the test, find the percentage of examined instances that had true positives and true negatives. According to the computations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision is the number of affirmative cases or the classification's accuracy rate. The following formula is applied to assess accuracy:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: A model's ability to recognise every instance of a pertinent machine learning class is measured by its recall. The ratio of accurately predicted positive observations to the total number of positives indicates how well a model can identify class instances.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: Mean Average Precision is one ranking quality metric (MAP). It considers the number of relevant recommendations and their position on the list. MAP at K is calculated as the arithmetic mean of the Average Precision (AP) at K for each user or query.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: An accurate machine learning model is indicated by a high F1 score. combining precision and recall to increase model correctness. The accuracy statistic quantifies the frequency with which a model correctly predicts a dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

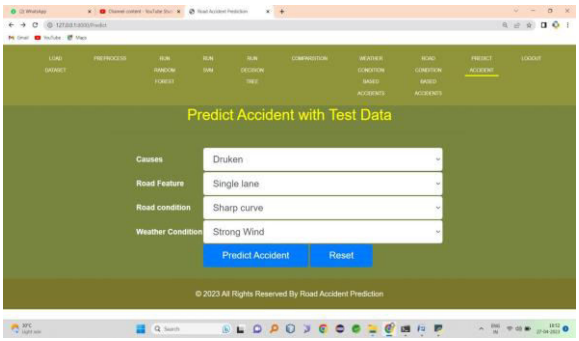

Fig2 input data

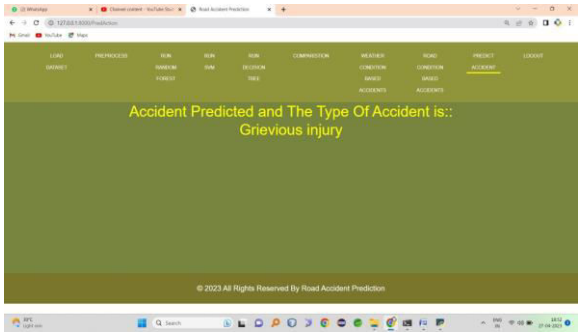

Fig3 results

5. CONCLUSION

This paper presented a hybrid machine learning-based road accident prediction system that integrates data preprocessing, feature selection, clustering, and classification techniques. By utilizing algorithms such as Support Vector Machine (SVM), Decision Tree, and Random Forest, the system effectively analyzes historical accident data and predicts the likelihood of accident occurrences. The combination of clustering with classification improves the accuracy and reliability of predictions.

The experimental results demonstrate that the proposed model can successfully identify accident-prone conditions and assist in making informed decisions to reduce road accidents. This system can be beneficial for government authorities, traffic management systems, and road users by providing early warnings and enhancing road safety measures.

6. FUTURE SCOPE

The proposed system can be further enhanced by integrating real-time data sources such as IoT sensors, GPS, and live traffic information to improve prediction accuracy and timeliness. Incorporating deep learning techniques and advanced models can also help in capturing more complex patterns in accident data. Additionally, the system can be deployed as a real-time web or mobile application to provide instant alerts to drivers and authorities.

Future improvements may include expanding the dataset with more diverse and large-scale data, including driver behavior and vehicle conditions. Integration with smart city infrastructure and traffic control systems can further enhance its practical applicability, enabling proactive accident prevention and efficient traffic management.

REFERENCES

[1] https://www.statista.com/topics/5982/road-accidents-in-india/

[2] Srivastava AN, Zane-Ulman B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In Aerospace Conference, IEEE. IEEE 3853-3862.

[3] Ghazizadeh M, McDonald AD, Lee JD. (2014). Text mining to decipher free-response consumer complaints: Insights from the nhtsa vehicle owner's complaint database. Human Factors 56(6): 1189-1203. http://dx.doi.org/10.1504/IJFCM.2017.089439.

[4] Chen ZY, Chen CC. (2015). Identifying the stances of topic persons using a model-based expectationmaximization method. J. Inf. Sci. Eng 31(2): 573-595. http://dx.doi.org/10.1504/IJASM.2015.068609.

[5] Williams T, Betak J, Findley B. (2016). Text mining analysis of railroad accident investigation reports. In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009-V001T06A009. http://dx.doi.org/10.14299/ijser.2013.01.

[6] Suganya, E. and S. Vijayarani. "Analysis of road accidents in India using data mining classification

algorithms." 2017 International Conference on Inventive Computing and Informatics (ICICI) (2017): 1122-1126.

[7] Sarkar S, Pateshwari V, Maiti J. (2017). Predictive model for incident occurrences in steel plant in India. In ICCCNT 2017, IEEE, pp. 1-5. http://dx.doi.org/10.14299/ijser.2013.01.

[8] Stewart M, Liu W, Cardell-Oliver R, Griffin M. (2017). An interactive web-based toolset for knowledge discovery from short text log data. In

International Conference on Advanced Data Mining and Applications. Springer, pp. 853-858. http://dx.doi.org/10.1007/978-3- 319-69179-

4_61.

[9] Zheng CT, Liu C, Wong HS. (2018). Corpus based topic diffusion for short text clustering. Neurocomputing 275: 2444-2458.

http://dx.doi.org/10.1504/IJIT.2018.090859.

[10] ArunPrasath, N and Muthusamy Punithavalli. "A review on road accident detection using data mining techniques." International Journal of Advanced Research in Computer Science 9 (2018): 881-885.

[11] George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Thomas Richter, Stephan Ruhl, Francesca La Torre, Lorenzo Domenichini, Daniel Graham, Niovi Karathodorou, Haojie Li (2016). "Use of

accident prediction models in road safety management – an international inquiry". Transportation Research Procedia 14, pp. 4257

– 4266.

[12] Anand, J. V. "A Methodology of Atmospheric Deterioration Forecasting and Evaluation through Data Mining and Business Intelligence." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 02 (2020): 79-87.

[13] Prayag Tiwari, Sachin Kumar, Denis Kalitin (2017). "Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques". International Conference on Computational Intelligence, Communications, and Business Analytics. 10.1007/978-981-10-6430-2_31.

[14] Kaur, G. and Er. Harpreet Kaur. "Prediction of the cause of accident and accident prone location on roads using data mining techniques."

2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (2017): 1-7.

[15] Irina Makarova, Ksenia Shubenkova, Eduard Mukhametdinov, and Anton Pashkevich, "Modeling as a Method to Improve Road Safety

During Mass Events", Transportation Research Procedia 20 (2017) 43.

Author Profiles

Mrs.K.Pavani is working as an Assistant and Head of Department of MCA, in SRK Institute of technology in Vijayawada. She completed her MCA and M.Tech in Computer Science. She has 10 years of teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her areas of interest include AI and ML, etc.

Ms.A.Anjali is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc.(computers) from Sri Sriddhartha Degree College Nuzvid. Her area of interest are DBMS and Machine Learning with Python.